Accuracy Optimization in Speech Pathology Diagnosis with Data Preprocessing Techniques

Joana Filipa Teixeira Fernandes[1,2(✉)] ⓘ, Diamantino Rui Freitas[2] ⓘ,
and João Paulo Teixeira[1,3] ⓘ

[1] Research Centre in Digitalization and Intelligent Robotics (CeDRI), Instituto Politécnico de
Bragança (IPB), 5300 Bragança, Portugal
{joana.fernandes,joaopt}@ipb.pt
[2] Faculty of Engineering, University of Porto (FEUP), 4200-465 Porto, Portugal
dfreitas@fe.up.pt
[3] Associate Laboratory for Sustainability and Technology (SusTEC), Instituto Politécnico de
Bragança, Campus de Santa Apolónia, 5300-253 Bragança, Portugal

Abstract. Using acoustic analysis to classify and identify speech disorders non-invasively can reduce waiting times for patients and specialists while also increasing the accuracy of diagnoses. In order to identify models to use in a vocal disease diagnosis system, we want to know which models have higher success rates in distinguishing between healthy and pathological sounds. For this purpose, 708 diseased people spread throughout 19 pathologies, and 194 control people were used. There are nine sound files per subject, three vowels in three tones, for each subject. From each sound file, 13 parameters were extracted. For the classification of healthy/pathological individuals, a variety of classifiers based on Machine Learning models were used, including decision trees, discriminant analyses, logistic regression classifiers, naive Bayes classifiers, support vector machines, classifiers of closely related variables, ensemble classifiers and artificial neural network classifiers. For each patient, 118 parameters were used initially. The first analysis aimed to find the best classifier, thus obtaining an accuracy of 81.3% for the Ensemble Sub-space Discriminant classifier. The second and third analyses aimed to improve ground accuracy using preprocessing methodologies. Therefore, in the second analysis, the PCA technique was used, with an accuracy of 80.2%. The third analysis combined several outlier treatment models with several data normalization models and, in general, accuracy improved, obtaining the best accuracy (82.9%) with the combination of the Greebs model for outliers treatment and the range model for the normalization of data procedure.

Keywords: Outliers · Normalization · Speech Pathologies · Speech Features · Machine Learning · Vocal Acoustic Analysis

1 Introduction

This research aims to develop a straightforward artificial intelligence model that can distinguish between healthy and pathological subjects with high accuracy rates and can be implemented in a system for the early detection of vocal pathologies. In its initial

trial stage, this technology will be installed in hospitals, where it will be used to record people's voices and determine if they are healthy or pathological.

Some unique data points differ significantly from other observations in a dataset. These observations are called outliers. Finding these outlying/anomalous observations in datasets has recently attracted much attention and is significant in many applications [1, 2].

As a rule, the appearance of outliers in databases is mainly due to human errors, instrument errors, population deviation, fraudulent behaviour and changes or failures in the system's behaviour. Some outliers can be observed as natural data points [3]. Detecting outliers in a dataset is important for many applications, such as network analysis, medical diagnostics, agricultural intelligence and financial fraud detection [4]. The statistics-based outlier detection methods model the objects using mean and standard deviation for a Gaussian distribution dataset or the median and inter-quartile range for non-Gaussian distribution [5–8].

Because normalization operations are designed to reduce issues like data redundancy and skewed results in the presence of anomalies, several modelling techniques, like Neural Networks, KNN, and clustering, benefit from improved performance [9].

The data set underwent some modifications. Making a scale to stabilize variance, lessen asymmetry, and bring the variable closer to the normal distribution is therefore what is needed [9].

In sets of searches with excessive information due to the inclusion of many features, issues like high dimensionality, overfitting risk, and biased results are dealt with by the selection of features. Low relevance and redundant input data have an impact on learning algorithms [10, 11].

As a result, the initial data set's size is reduced, computational costs are decreased, and the forecast accuracy of the predictors is increased as a result of the selection and elimination of less important attributes. The search direction, the search methodology, and the stopping criterion are the three dimensions that make up the selection of features [12].

This work intends to find a classification model and optimize the accuracy in the classification between healthy and pathological subjects. Therefore, it is necessary to treat and correct the anomalies identified in the automatically extracted feature values available in systems related to the diagnosis of voice pathologies, which have as input parameters relative jitter, absolute jitter, RAP jitter, PPQ5 jitter, absolute shimmer, relative shimmer, APQ3 shimmer, APQ5 shimmer, Fundamental Frequency, Harmonic to Noise Ratio (HNR), autocorrelation, Shannon entropy, logarithmic entropy and the subject's sex [13–17].

This work includes 4 section, the first being the introduction. The second describes the database, extracted parameters, outlier identification and treatment methods, data normalization and main component analysis. In the third chapter, the results and discussion are described. Lastly, the conclusion.

2   Materials and Methods

In this section, the Database used, the parameters used, the outliers identification methods, the normalization methods, Principal Component Analysis (PCA) and evaluation measurements will be described.

2.1   Database

The German Saarbrucken Voice Database (SVD), made available online by the Institute of Phonetics at the University of Saarland, was used as the source for the speech files [18].

The collection includes voice signals from more than 2000 individuals with both healthy voices and vocal problems. Each subject has recordings of the German greeting "Good morning, how are you?" with the phonemes /a/, /i/, and /u/ in the low, neutral/normal, and high tones and shifting between tones. The sound files have a duration of between one and three seconds, were recorded in mono at a sample rate of 50 kHz, and have a resolution of 16 bits [19]. Table 1 includes the mean age and standard deviation as well as the distribution of subjects by various pathologies (19 diseases, a total of 708 diseased subjects, and 194 control subjects).

2.2   Feature Extraction

In this section, the various parameters that will be extracted from the speech signal will be described.

Jitter is the glottal fluctuation between vocal cord vibration cycles. Higher jitter values are typically observed in subjects who have trouble modulating their vocal cords. Absolute Jitter (jitta), Relative Jitter (jitter), Relative Average Perturbation Jitter (RAP), and Five-point Period Perturbation Quotients Jitter (PPQ5) will be used as input features [15, 17, 19].

Shimmer is the amplitude variation over the glottal periods. Variations in glottal magnitude are mostly caused by lesions and decreased glottal resistance. Reduced glottal resistance and injuries can result in higher shimmer values, which can change the glottal magnitude. Absolute Shimmer (ShdB), Relative Shimmer (shim), Three-Point Amplitude Perturbation Quotient Shimmer (APQ3), and Five-Point Amplitude Perturbation Quotient Shimmer (APQ5) will be used as the shimmer measurements [15, 17, 19].

Fundamental Frequency (F0) is thought to correspond to the vibration frequency of the vocal cords. The Autocorrelation method is used to calculate the F0, with a frame window length of 100 ms and a minimum F0 of 50 Hz [20].

Harmonic to Noise Ratio (HNR) enables for assessing the relationship between the harmonic and noise components of a speech signal. Different vocal tract topologies result in various amplitudes for the harmonics, which can cause the HNR value of a signal to change [16, 21–24].

Table 1. Groups used for the study, sample size, mean and standard deviation of the ages.

Groups		Sample size		Average Ages	Standard Deviation Age
		subjects	%		
Control		194	21.51	38.06	14.36
Pathological	Dysphonia	69	7.65	47.38	16.27
	Chronic Laryngitis	41	4.55	49.69	13.47
	Vocal Cord Paralysis	169	18.74	57.75	13.77
	Cyst	3	0.33	47.50	15.56
	Vocal Strings Polyp	27	2.99	52.28	13.41
	Carcinoma of Vocal Strings	19	2.11	57.00	6.60
	Laryngeal Tumor	4	0.44	53.50	8.17
	Granuloma	2	0.22	44.50	4.50
	Intubation granuloma	3	0.33	53.00	11.22
	Hypopharyngeal Tumor	6	0.67	59.50	9.29
	Fibroma	1	0.11	46.00	0.00
	Dysplastic Larynx	1	0.11	69.00	0.00
	Reinke's edema	34	3.77	56.10	11.37
	Functional Dysphonia	75	8.31	47.12	14.54
	Hypofunctional Dysphonia	12	1.33	41.63	15.07
	Hyperfunctional Dysphonia	127	14.08	42.32	13.62
	Hypotonic Dysphonia	2	0.22	49.50	12.50
	Psychogenic Dysphonia	51	5.65	51.40	9.40
	Spasmodic Dysphonia	62	6.87	57.15	15.75

Autocorrelation.

The autocorrelation gives an indication of how similar succeeding phonatory periods that are repeated throughout the signal are to one another. The signal's periodicity increases as the autocorrelation value rises [19, 21, 25].

Entropy.

In order to quantitatively quantify the level of unpredictability and uncertainty of a particular data sequence, it takes into account the energy that is present in a complex system. Entropy analysis makes it feasible to precisely evaluate the nonlinear behavior characteristic of voice signals [26].

2.3   Identification and Treatment of Outliers

The basic methods for finding outliers can be distinguished by the criteria used, such as classification, distance, densities, clusters, and statistics [3].

The calculation of mean, standard deviation and histograms is affected by outliers. As a result, it distorts generalizations and inferences about the studied data set. As a result, the inclusion of outliers in the dataset can result in incorrect interpretations [3].

In the *Median* method, Outliers are items that deviate more than three MED from the median. Equation 1 provides the MED scale's definition.

$$c * \text{median}(|A - \text{median}(A)|) \tag{1}$$

where A is the data and c is described by Eq. 2, where erfcinv is the inverse complementary error function [27].

$$c = \frac{-1}{\sqrt{2} * \text{erfcinv}\left(\frac{3}{2}\right)} \tag{2}$$

The *Mean* method defines Outliers by the Mean method as components that deviate from the mean by more than three standard deviations. This approach is quicker but less reliable than the median approach [8, 27, 28].

In the *Quartile* method, items with more than 1.5 inter-quartile range above the upper quartile (75%) or below the lower quartile (25%), are considered outliers. This approach is advantageous when the data has not a normal distribution [27, 28].

The Grubbs test, which eliminates one outlier per iteration based on the hypothesis test, is used by the *Grubbs* method to identify outliers. The data will be assumed to have a normal distribution for this method [27].

By employing the Grubbs test, which eliminates one outlier per iteration based on the hypothesis test, the *Gesd* technique finds outliers. According to this strategy, the data should have normal distribution [27].

Once an outlier has been identified, filling is the procedure used to handle it. The limit value, which is determined in accordance with the selected method, takes the place of the outlier.

When new subjects (samples) are included in the dataset, the recognition process must be verified using the threshold value that was previously established using the original data set.

2.4 Normalization

Some modeling tools, such as neural networks, the k-nearest neighbors algorithm (KNN), and clustering, benefit from normalization since these normalizing operations aim to reduce issues such data redundancy and skewed findings in the presence of anomalies. The dataset underwent several modifications. Therefore, the goal is to create a scale that will stabilize a variance, reduce asymmetry, and approach the variable's normal distribution [3].

The ***Z-Score*** calculates a data point's distance from the mean, related to standard deviation,. The original data set's shape characteristics are preserved in the standardized data set (same skewness and kurtosis), which has a mean of 0 and a standard deviation of 1 [29, 30].

The general definition of the P-norm of a vector v with N elements according to the ***P-Norm*** technique is: where p is any positive real value, Inf or -Inf. Typical values of p include 1, 2, and Inf [29, 30].

- The sum of the absolute values of the vector elements is the 1-norm that results if p is 1.
- The vector magnitude, or Euclidean length of the vector, is determined by the 2-norm that results if p is 2.
- When p is Inf, then $\|v\|_\infty = max_i(|v(i)|)$.

By stretching or compressing the points along the number line, the ***Resizing*** method modifies the distance between the minimum and maximum values in a data collection. The data's z-scores are kept, therefore the statistical distribution's form is unaltered. The formula for scaling data X to a range [a, b] is: If A is constant, normalize returns the interval's lower limit (which is 0 by default) or NaN (when the range contains Inf) [29, 30].

A data set's ***Interquartile Range*** (IQR) describes the range of the middle 50% of values after sorting the values. In this case, the median of the data would be Q2, the median of the lower half would be Q1, and the median of the upper half would be Q3. When the data contains outliers (extremely big or very tiny values), the IQR is typically favored over examining the entire range of the data because it excludes the largest 25% and smallest 25% of values in the data [30].

The median value of the absolute deviations from the median of the data is known as the ***Median Absolute Deviation*** (MAD) of a data collection. As a result, the MAD illustrates how variable the data are in regard to the median. When the data contains outliers (extremely big or very tiny values), the MAD is typically favored over using the standard deviation of the data since the standard deviation squares differences from the mean, giving outliers an excessively significant impact. In contrast, the MAD value is unaffected by the deviations of a few outliers [30].

2.5 Principal Component Analysis (PCA)

This technique uses mathematical concepts such as standard deviation, covariance of eigenvalues and eigenvector. To determine the number of principal components, eigenvectors and eigenvalues must be determined starting from the covariance matrix. Then,

calculating the cumulative proportion of the eigenvalues is all that is required. As a result, the first eigenvectors that correspond to 90% or 95% of the collected percentage will be chosen, meaning that the first eigenvectors account for 90% or 95% of the data. The final step is to multiply the fitted data by the inverse of the chosen eigenvector matrix [31].

2.6 Evaluation Measurements

In order to evaluate the performance, accuracy will be used. This measure is observed in Eq. 3. However, the data used are unbalanced, hence the need to present 4 measures in addition to accuracy, namely precision, sensitivity, specificity and F1-score. These measures are presented in Eq. 4, 5, 6 and 7 respectively. Where TP stands for True Positive, FN stands for False Negative, FP is False Positive, TN is True Negative, P stands Precision and S is Sensibility.

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{3}$$

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$Sensibility = \frac{TP}{TP + FN} \tag{5}$$

$$Specificity = \frac{TN}{TN + FP} \tag{6}$$

$$F1 - score = 2 \times \frac{P \times S}{P + S} \tag{7}$$

3 Results and Discussion

In this chapter the results are presented as well as a discussion about them.

3.1 Results

For the analysis, 9 sound files were used per subject, and 13 parameters were extracted from each file (relative jitter, absolute jitter, RAP jitter, PPQ5 jitter, absolute shimmer, relative shimmer, APQ3 shimmer, APQ5 shimmer, fundamental frequency, HNR, autocorrelation, Shannon entropy and logarithmic entropy), giving 117 parameters per subject, to which sex was added. Therefore, the input matrix is composed of 118 lines × N number of subjects.

Having the input matrix, the classification between healthy and pathological began. As classifiers we used Decision Trees, Discriminant Analysis, Logistic Regression Classifiers, Naive Bayes Classifiers, Support Vector Machines, Nearest Neighbor Classifiers, Ensemble Classifiers and Artificial Neural Network. The cross-validation technique of 10 folds was applied during the training process.

The classifier that obtained the best result without any data pre-processing, with a binary output (control/pathological) was the Ensemble Subspace Discriminant [32] with an accuracy of 81.3%. This model had 30 learners and subspace dimension 59.

In order to improve this accuracy, the technique of reducing the dimension was used, using Principal Component Analysis (PCA). This analysis was applied to the 118 parameters, with a variance of 95%, resulting in 7 new features and an accuracy of 80.2% was obtained.

Given that the accuracy obtained with the PCA technique is lower than those obtained without any feature dimension reduction and considering the work of Silva et al. 2019 [3], where it obtained an improvement of up to 13 percentage points, an attempt was made to understand whether, with the treatment of outliers and data normalization, the accuracy of the classifier increased. Therefore, in Table 2 it is possible to observe the result of the various combinations between the various models for treating outliers with the various models for normalizing the data. In this analysis, PCA was not used, since there was a loss of accuracy. In the normalization using the range model (resizing method), the data were normalized between $[-1, 1]$.

Table 2. Accuracy using the Ensemble Subspace Discriminant model with the various outlier detection and data normalization techniques.

Accuracy (%)		Normalization method					
		zscore	norm	scale	range	center	medianiqr
Outliers method	Median	82.3	82	82.3	81.6	81.6	82.4
	Mean	81.6	81.2	82	81.9	82.2	81.5
	Quartiles	82.3	82.0	82.4	82.5	82.5	82.5
	Grubbs	82.5	81.7	82.3	**82.9**	82.3	82.2
	Gesd	82.2	82.0	81.2	81.8	82.3	81.6

Table 2 shows that using the outlier identification method and data normalization allows for improvements in accuracy over the baseline accuracy of 81.3%.

The combination between the outlier identification method and the data normalization method that obtained the best results was with the Grubbs method for outliers and the range method for data normalization, which obtained an accuracy of 82.9%. In this way, an improvement of 1.6 percentage points was achieved compared to the result where there was no data pre-processing, and an improvement of 2.7 percentage points compared to the accuracy obtained by the PCA method.

In Fig. 1, it is possible to see that the Area Under Curve (AUC) improved considerably, changing from 0.78 to 0.82.
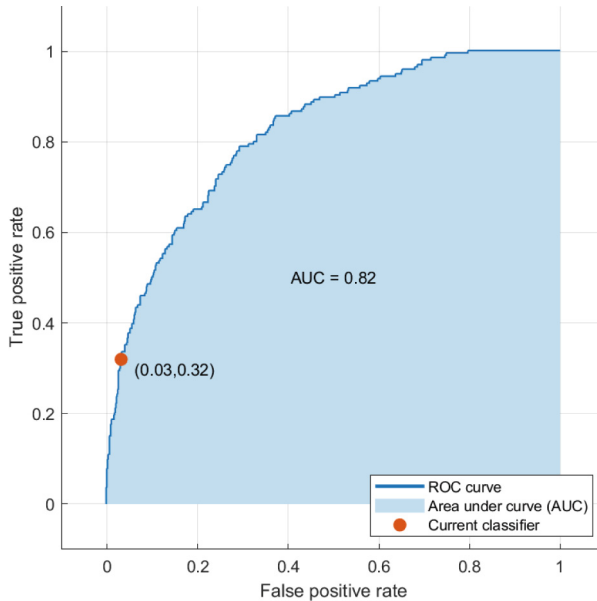
3.2 Discussion

Comparing the results obtained in this work with those obtained by Silva et al. 2019 [3], it is possible to notice that the results obtained in this work are similar. In both works,
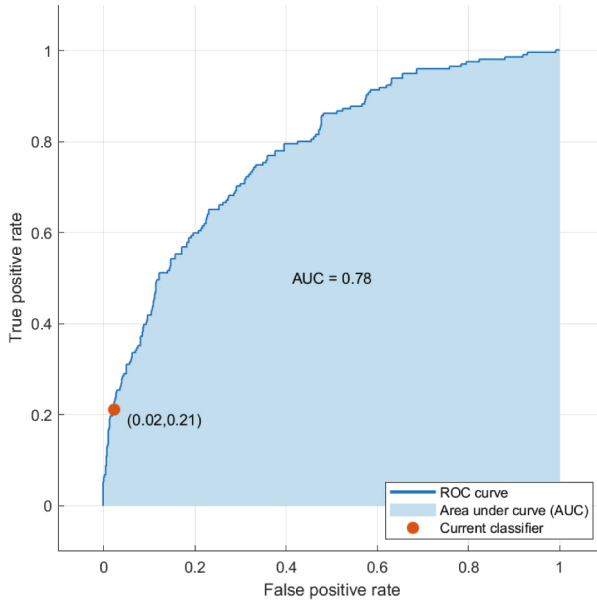
a)   Classifier ROC curve without data pre-processing.

b) ROC curve of the classifier using the outliers and normalization method with better accuracy.

Fig. 1. a) Classifier ROC curve without data pre-processing; b) ROC curve of the classifier using the outliers and normalization method with better accuracy.

an improvement in the classification was obtained. However, in the work developed by Silva et al. 2019, the results obtained with the treatment of outliers showed a greater improvement at a percentage level between the results without processing outliers and those obtained with the treatment of outliers. This can be justified by the classifier used, since different classifiers are used and without any treatment of the data, in this work a higher accuracy was obtained, as well as, by the fact that in this work more pathologies are used, which leads to a great diversification of data, while in the work by Silva et al. 2019 [3] try to classify only between control and dysphonia, control and laryngitis and control and vocal cord paralysis and the data difference between control and pathology is smaller, that is, the data are not unbalanced. Besides, the baseline accuracy used in [3] was lower, between 63% and 80%, for different classification cases, leaving more space for improvements. Also, the methods of identification and treatment of outliers and data normalization used in this work and [3] are different. In Silva et al. 2019 [3] used the boxplot method and the standard deviation method as a method of identifying and treating outliers, and the z-score, logarithmic and square root method as a data normalization method.

For the situation with the best accuracy, precision, sensitivity, specificity and F1-score were calculated, obtaining 73.8%, 32%, 96.9% and 44.6% respectively.

The F1-score value is significantly different from the accuracy value, since the dataset is not balanced.

4   Conclusion

In order to try to obtain greater accuracy in the subject classification process (healthy/pathological) we tried to understand whether the results were better with the data from the input matrix in 3 ways: without any pre-processing, with PCA and without PCA with technique of identification and treatment of outliers and data normalization. Therefore, for this analysis we had 708 sick participants and 194 control individuals were used in this study, since it took into account 19 different pathologies.

Each subject comprises nine sound files, corresponding to three vowels and three tones, where 13 parameters were taken from each sound file, totalling 117 input features for each subject, to which the subject's sex was also added. The input matrix is thus made up of N subjects x 118 lines.

In this work, a first classification was started where the input matrix did not have any type of data pre-processing 8 types of classifiers with several models. The cross-validation technique of 10 validations was applied to these classifiers.From this first analysis, an accuracy of 81.3% was obtained for the Ensemble Subspace Discriminant classifier.

Then a second analysis was carried out where the Principal Component Analysis (PCA) technique was applied to the input matrix. In this analysis, only the classifier that obtained the best accuracy was used, but the accuracy results were not better, as an accuracy of 80.2% was obtained.

In work by Silva et al. 2019 [3], using different outlier treatment methods and data normalization, improved accuracy by up to 13 percentage points from a lower baseline accuracy. In this way, an analysis was initiated in which 5 outlier treatment models

were combined with 6 data normalization models without the use of PCA, for the same classification model. Therefore, an improvement of 1.6 percentage points was achieved, with an accuracy of 82.9%. This accuracy was obtained with the combination of the grubbs model in the treatment of outliers, with the range model in the normalization of the data.

As future work, it is intended to classify the types of signals. In signal classification there are 3 types of signals. In type 1 the signals are periodic, in type 2 they have some periodicity and in type 3 the signals are chaotic. Signals that are classified as type 3 cannot use these parameters since they are signals without any type of periodicity, which leads to extremely high jitter and shimmer values, thus impairing the classification. Later, it is intended to identify the pathology and the degree of severity.

In order to increase the database, this system is implemented in a hospital in order to collect more speech signals.

References

1. Toller, M.B., Geiger, B.C., Kern, R.: Cluster purging: efficient outlier detection based on rate-distortion theory. IEEE Trans. Knowl. Data Eng. **35**(2), 1270–1282 (2023). https://doi.org/10.1109/TKDE.2021.3103571
2. Abhaya, A., Patra, B.K.: An efficient method for autoencoder based outlier detection. Exp. Syst. Appl. **213**, 118904 (2023). https://doi.org/10.1016/J.ESWA.2022.118904
3. Silva, L., et al.: Outliers treatment to improve the recognition of voice pathologies. Procedia Comput. Sci. **164**, 678–685 (2019). https://doi.org/10.1016/J.PROCS.2019.12.235
4. Du, X., Zuo, E., Chu, Z., He, Z., Yu, J.: Fluctuation-based outlier detection. Sci. Rep. **13**(1), 2408 (2023). https://doi.org/10.1038/s41598-023-29549-1
5. Grubbs, F.E.: Procedures for detecting outlying observations in samples. Technometrics **11**(1), 1–21 (1969). https://doi.org/10.1080/00401706.1969.10490657
6. Atkinson, A.C., Hawkins, D.M.: Identification of outliers. Biometrics **37**(4), 860 (1981). https://doi.org/10.2307/2530182
7. Yang, X., Latecki, L.J., Pokrajac, D.: Outlier detection with globally optimal exemplar-based GMM. In: 2009 9th SIAM International Conference on Data Mining. Proceedings in Applied Mathematics, vol. 1, pp. 144–153. Society for Industrial and Applied Mathematics (2009). https://doi.org/10.1137/1.9781611972795.13
8. Seo, S., Marsh, P.D.G.M.: A review and comparison of methods for detecting outliersin univariate data sets (2006). http://d-scholarship.pitt.edu/7948/
9. Pino, F.A.: A questão da não normalidade: uma revisão. Rev. Econ. Agrícola **61**(2), 17–33 (2014)
10. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. J. Mach. Learn. Res. **3**, 1157–1182 (2003)
11. Rodrigues, P.M., Teixeira, J.P.: Classification of electroencephalogram signals using artificial neural networks. In: Proceedings of the 2010 3rd International Conference on Biomedical Engineering and Informatics, BMEI 2010, vol. 2, pp. 808–812 (2010). https://doi.org/10.1109/BMEI.2010.5639941

12. Silva, L., Bispo, B., Teixeira, J.P.: Features selection algorithms for classification of voice signals. Procedia Comput. Sci. **181**, 948–956 (2021). https://doi.org/10.1016/J.PROCS.2021.01.251

13. Teixeira, J.P., Freitas, D.: Segmental durations predicted with a neural network. In: International Conference on Spoken Language Processing, Proceedings of Eurospeech 2003, pp. 169–172 (2003)

14. Teixeira, J.P., Freitas, D., Braga, D., Barros, M.J., Latsch, V.: Phonetic events from the labeling the European Portuguese database for speech synthesis, FEUP/IPB-DB. In: International Conference on Spoken Language Processing, Proceedings of Eurospeech 2001, pp. 1707–1710 (2001). 8790834100, 978-879083410-4

15. Teixeira, J.P., Gonçalves, A.: Algorithm for jitter and shimmer measurement in pathologic voices. Procedia Comput. Sci. **100**, 271–279 (2016). https://doi.org/10.1016/J.PROCS.2016.09.155

16. Fernandes, J., Teixeira, F., Guedes, V., Junior, A., Teixeira, J.P.: Harmonic to noise ratio measurement - selection of window and length. Procedia Comput. Sci. **138**, 280–285 (2018). https://doi.org/10.1016/J.PROCS.2018.10.040

17. Fernandes, J., Junior, A.C., Freitas, D., Teixeira, J.P.: Smart data driven system for pathological voices classification. In: Pereira, A.I., Košir, A., Fernandes, F.P., Pacheco, M.F., Teixeira, J.P., Lopes, R.P. (eds.) Optimization, Learning Algorithms and Applications: Second International Conference, OL2A 2022, Póvoa de Varzim, Portugal, October 24–25, 2022, Proceedings, pp. 419–426. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-23236-7_29

18. Pützer, M., Barry, W.J.: Saarbruecken Voice Database. Institute of Phonetics at the University of Saarland (2007). http://www.stimmdatenbank.coli.uni-saarland.de. Accessed 05 Nov 2021

19. Fernandes, J., Silva, L., Teixeira, F., Guedes, V., Santos, J., Teixeira, J.P.: Parameters for vocal acoustic analysis - cured database. Procedia Comput. Sci. **164**, 654–661 (2019). https://doi.org/10.1016/J.PROCS.2019.12.232

20. Hamdi, R., Hajji, S., Cherif, A., Processing, S.: Recognition of pathological voices by human factor cepstral coefficients (HFCC). J. Comput. Sci. **16**, 1085–1099 (2020). https://doi.org/10.3844/jcssp.2020.1085.1099

21. Fernandes, J.F.T., Freitas, D., Junior, A.C., Teixeira, J.P.: Determination of harmonic parameters in pathological voices—efficient algorithm. Appl. Sci. **13**(4), 2333 (2023). https://doi.org/10.3390/app13042333

22. Teixeira, J.P., Fernandes, P.O.: Acoustic analysis of vocal dysphonia. Procedia Comput. Sci. **64**, 466–473 (2015). https://doi.org/10.1016/J.PROCS.2015.08.544

23. Teixeira, J.P., Fernandes, J., Teixeira, F., Fernandes, P.O.: Acoustic analysis of chronic laryngitis statistical analysis of sustained speech parameters. In: 11th International Joint Conference on Biomedical Engineering Systems and Technologies, BIOSTEC 2018, vol. 4, pp. 168–175 (2018). https://doi.org/10.5220/0006586301680175

24. Boersma, P.: Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In: IFA Proceedings 17, vol. 17, pp. 97–110 (1993). http://www.fon.hum.uva.nl/paul/papers/Proceedings_1993.pdf

25. Boersma, P.: Stemmen meten met Praat. Stem-, Spraak- en Taalpathologie **12**(4), 237–251 (2004)

26. Araújo, T., Teixeira, J.P., Rodrigues, P.M.: Smart-data-driven system for alzheimer disease detection through electroencephalographic signals. Bioengineering **9**(4), 141 (2022). https://doi.org/10.3390/bioengineering9040141

27. NIST/SEMATECH: e-Handbook of Statistical Methods. http://www.itl.nist.gov/div898/handbook/. Accessed 14 Jun 2023

28. Unwin, A.: Exploratory data analysis, 3rd edn. In: International Encyclopedia of Education, pp. 156–161. Elsevier, Amsterdam (2010). https://doi.org/10.1016/B978-0-08-044894-7.01327-0

29. Triola, M.F.: Introdução à estatística, 12th edn. In: Elementary Statistics. Pearson Education INC, Rio de Janeiro (2017)
30. MathWorks: Normalize. https://www.mathworks.com/help/matlab/ref/double.normalize.html#d124e1046230. Accessed 14 Jun 2023
31. Teixeira, J.P., Alves, N., Fernandes, P.O.: Vocal acoustic analysis: ANN Versos SVM in classification of dysphonic voices and vocal cords paralysis. Int. J. E-Health Med. Commun. **11**(1), 37–51 (2020). https://doi.org/10.4018/IJEHMC.2020010103
32. Ashour, A.S., Guo, Y., Hawas, A.R., Guan, Xu.: Ensemble of subspace discriminant classifiers for schistosomal liver fibrosis staging in mice microscopic images. Health Inf. Sci. Syst. **6**(1), 21 (2018). https://doi.org/10.1007/s13755-018-0059-8